Adriana Rakshana
Student ID: 10456351
Computational Content Analysis
University of Chicago Spring 2020

Two and a half men[1] - Gender disparity, not the TV show

Abstract

Conventional wisdom in Hollywood is that male stars are a bigger box office draw, often the reason given for their higher salaries. However, recently, it has been revealed that female-led films actually do outperform male ones at the box-office[2] This paper will build on such work highlighting diversity and inclusion in film; I hope to show how genders are represented within film, extracting common themes and word relations as they correspond to men and women. This will help reveal whether there are any inherent differences between male and female terminology over time and within genres. We will attempt to do this using both primary content analysis of around 10000 movie scripts from 1930 to 2017 [3]. We will do this by outlining some general differences between specific genres through word embedding methods as well as outline different semantic and parts of speech networks and vector space relations. Through these methods, we essentially find that there are actually specific ways in which women and men are described within the different films. The differences are more prominent over time comparing decades than within genre; earlier films had women described in a more submissive manner which becomes more explicit in the more recent years. We also find that certain genres of movies are more likely to be associated with women versus men. Men are more likely to be associated with war films and women are more likely to be associated with family films and musicals apparently.

[1] Referencing the around two and a half to one male to female ration in film. I could have perhaps thought of a better name.

[2] "Women-Led Films Dominate at the Box Office, Study Finds." The Guardian. Guardian News and Media, December 12, 2018. https://www.theguardian.com/film/2018/dec/12/women-outperform-men-at-the-box-office-study.

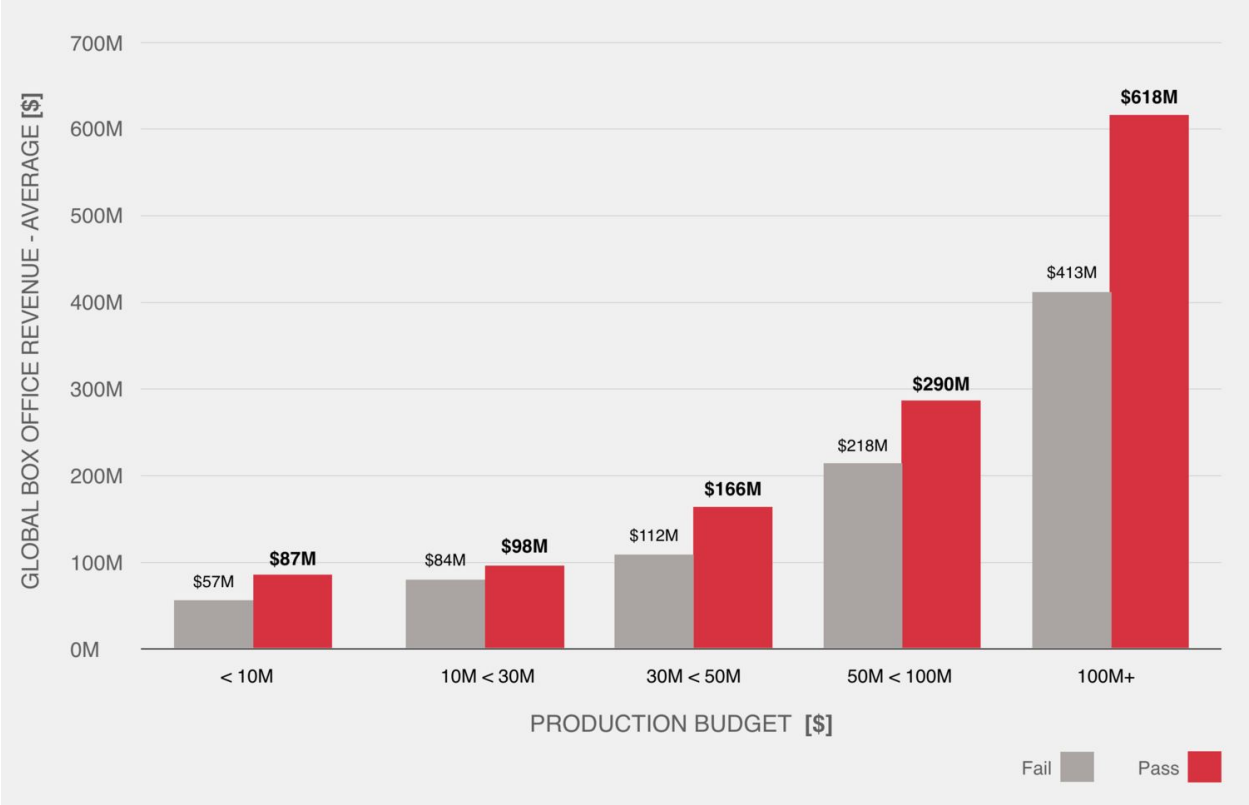
[3] I initially started of thinking I would research what would make a movie successful as far as gross revenue was concerned and did some research and quickly realized that there are a lot more factors to be concerned for movie success such as the actors and initial budget for example that go way beyond the impact of the themes and genre of the the script. Thus, we stick to the word related relations within the scripts.

Literature Review

Movies are often described as having the power to influence individual beliefs and values. In (Cape, 2003), the authors assert movies' influence in both creating new thinking patterns in previously unexplored social phenomena, especially in children, as well as their ability to update an individual's existing social boundaries based on what is shown on screen as the "norm". Thus, it is critical to study the way different characters have influence within films and alter this balance if it too skewed in one direction or another.

Previous works in studying representation in movies largely focus on relative frequencies, particularly on character gender as well as the box office revenue to some extent. According to research done on top-grossing films from 2014 to 2017, on average, female-led[4] films lead global box office revenue at every budget level for 2014-2017. Moreover, films that passed the Bechdel test - where two female characters have a conversation about something other than a man - made more revenue at the box office at every budget level than films that failed the test. Yes, the Bechdel test is quite a primitive measure of gender diversity within film, but far too many films do not even pass this low bar of inclusivity.

[4] The researchers define a "female lead" as a woman who is listed first in official press materials.



Automated analyses of movies using computational techniques to analyze representation has recently gained some attention. In (NYFA, 2013; Polygraph, 2016), the authors examine differences in relative frequency of female characters and note considerable disparities in gender ratio in these movies[5]. However, the analyses there too are limited to comparing relative frequencies. My work is closest to work done by the SAIL lab where the authors analyzed content of characters' language and their interactions across gender, race and age. Beyond the cast, the researchers also looked at genre, the production teams across films including writers, directors and casting agents. In my research, I will be focusing just on the

[5] Gálvez, Ramiro H., Valeria Tiffenberg, and Edgar Altszyler. "Half a century of stereotyping associations between gender and intellectual ability in films." Sex Roles 81, no. 9-10 (2019): 643-654.

computational content analysis of the scripts itself, not an analysis of the gender diversity on the backend of production.

Data

The data used for this project is the movie script corpus from the Davies corpora for the years 1930 to 2017. The number of scripts used totalled 9909 in order to allow for enough number of movies within each genre and decade . This will allow for enough movies within each category for content and time based comparisons about how women are written about in the movie scripts. If a smaller sample is used for the sake of computational restrictions, a note is made next to the analysis.

Data Analysis

The purpose of this study initially was to analyze films according to major organizing principles of society: gender, race, age, and sexual orientation. Given computational constraints, I restricted the analysis to gender. Thus, we examine how gender representation can be extracted from films by both time and genre.

Semantic Networks of all the genres

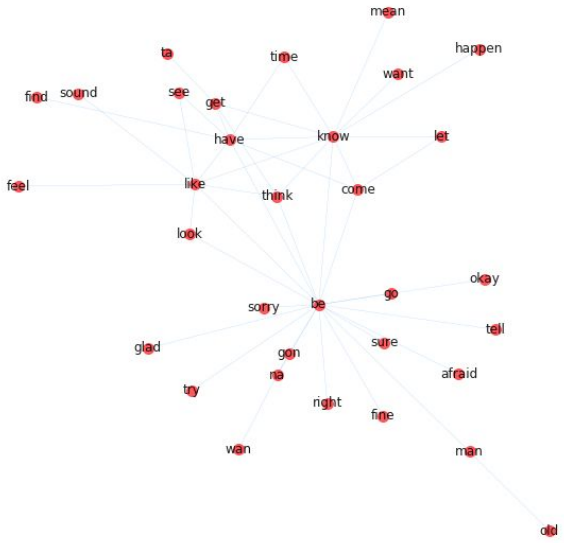
Before delving into the gender level representations, we can look at some general genre level semantic networks, to see what kind of word centralities exist to see what themes are commonly associated with each genre. I did this first by obtaining the tokenized and normalized sentences for each group of genres.

For this, I first did a general broad network of a certain subset of the whole corpus - romance, comedy, horror, drama, sci-fi and a mix of all the subsetted movies. Then, I used the word co-occurrence function with a sparse matrix due to memory issues to draw the different networks for the genres, however, these networks were quite crowded and did not really offer me much information in terms of genre identification. We then convert the collection of scripts to a matrix of token counts using the sk.learn vectorizer and look into the document to document relations . We remove the nodes with weights less than 50 since there were a lot mor enodes and edges for the romance movies. For our other genres, namely, sci-fi, horror and drama we just eliminated nodes less than 25 in weight. From the connected component subgraphs, we can see that the words tend to be quite similar for the different genre, with the most common central nodes being words like ‘be’, ‘know’, ‘come’ and ‘let’. These words are also the main words that show up in our word cloud associations for all the movies within the corpus. We can see words like love, marriage, man and feel quite central within the romance movies. However, what gave me the most distinctive

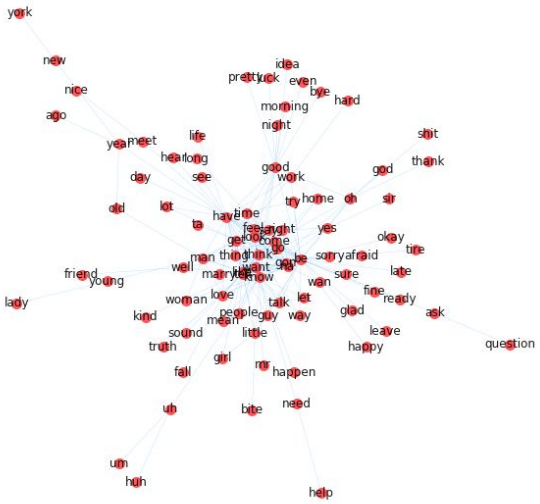
genre level centralities was actually the parts of speech level associations for the different genres which we will discuss next.

First level connected component subgraph

Sci-fi movies



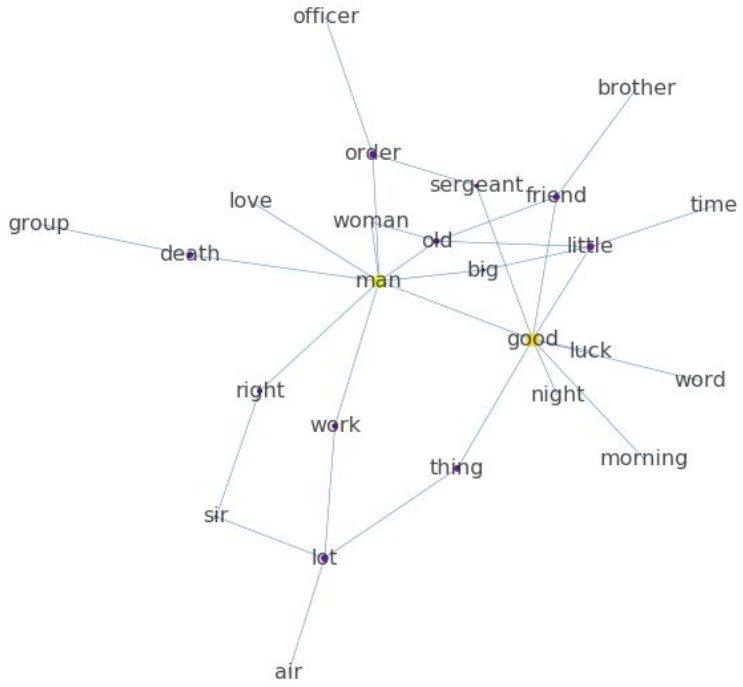
Romance movies



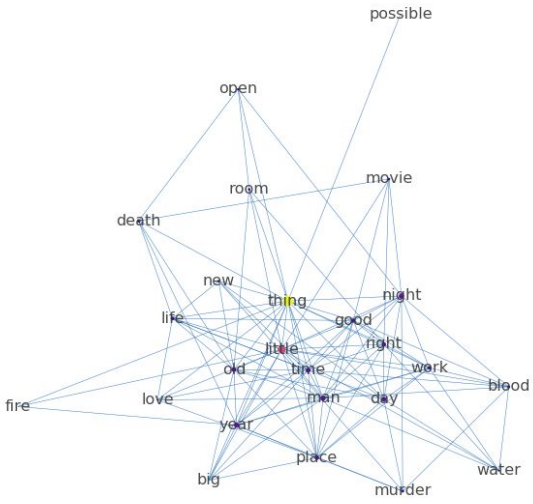
Parts of Speech Associations

When calculating the centralities for the different genres, I compared the betweenness centrality and the eigenvector centrality given that both the options seem the most relevant for our analysis. The betweenness centrality distinguishes nodes that require the most shortest pathways between all other nodes in the network which implies that words with a high betweenness centrality may link distinctive domains, rather than being "central" to any one. On the other hand, the eigenvector centrality weights the degree by the centrality of those to whom one is tied (and the degree to whom they are tied, etc.), i.e. an nth order degree measure. The betweenness centrality was the best option for my analysis in my opinion since it would actually link the distinctive domains which in my case would be the links between the genders within the genres and time period breakdowns and allow us to know about the amount of influence a node has over the flow of information. Closeness centrality wouldn't be really helpful since it is more of a measure of how long it will take to spread information from one to all other nodes sequentially, which isn't really what we're looking for when we want to establish representation. As can be seen below, these part of speech associations, particularly the noun-adjective relations. All the different genres have very distinct words that are revealed in the nodes that are quite specific to their genre with war movies having words like order, sergeant and officer showing up. Similarly, words like death, blood and murder take center stage for horror films. Based on these relevant parts of speech revelations for general genre level breakdowns, I further went on to query these relationships for gender specifically.

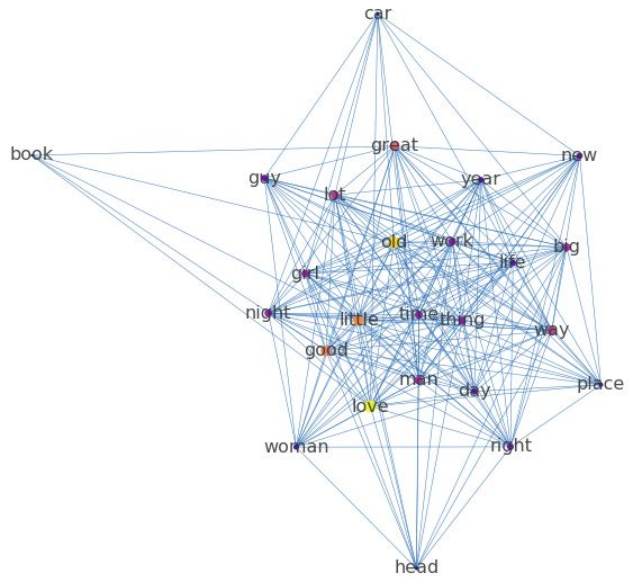
War



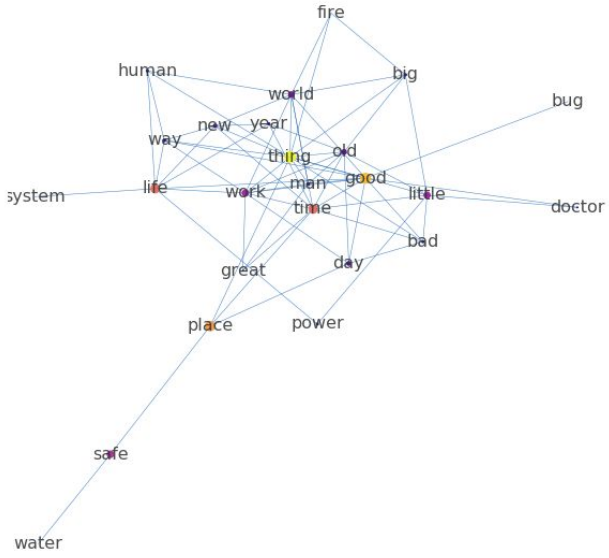
Horror



Romance



Sci-fi



POS for gender specifically

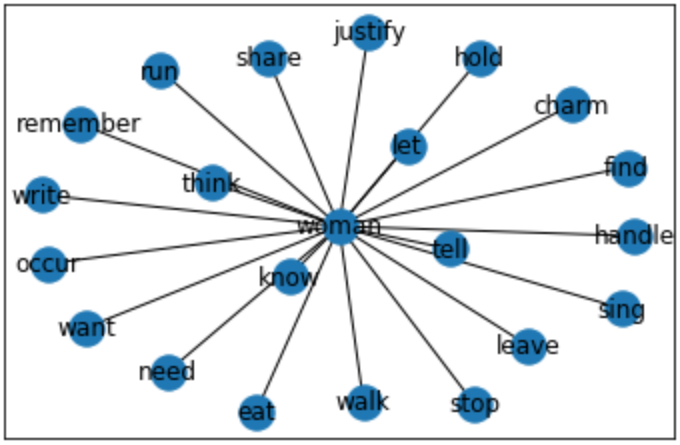
Given the parts of speech relations were quite distinctive above, we look at the time related distinctions for noun-verb and noun-adjectives to see how men and women are situated

within their respective networks between movies from the 1930s and movies from the 2010s. The first layer of nodes attached to both and women are largely dominated by what are the most common verbs for the corpus regardless of gender with words like let, think, come, tell and know so we query the second layer for women with the second and third layer for men. The fact that men have many more edges and node connections for verbs imply that men are more likely to be portrayed as those with agency responsible for different actions more so than women. This would align with the fact that we found that men had more dialogue than women in film over time. In addition to having lesser actions related to them, women's verb associations for the 1930s have terms like charm, want, need contrasted with words like grow, believe, expect, live, feel, fight and decide in addition to the words seen for women which suggests to me that men are written with more diverse emotions and character development and autonomy in terms of their actions. I might not be characterizing these differences adequately since I do not have much of a linguistic or cinema-studies background but it does still feel quite evident that women and men have different actions associated with them, or at least one gender clearly has a lot more **action** associated with their gender.

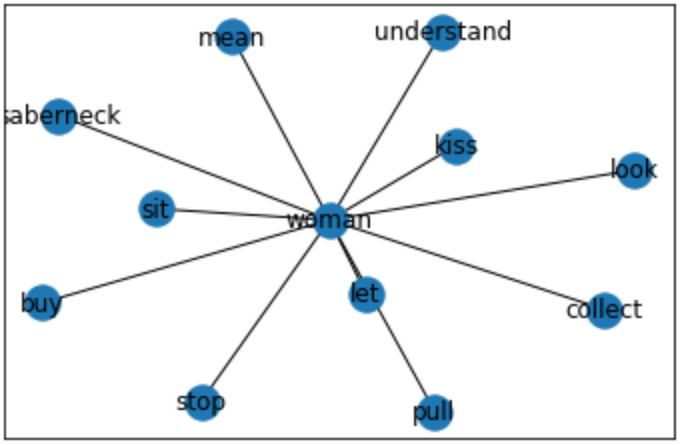
We can compare the verb associations from the 1930s to the 2010s. We remove any nodes with n1 and n2 bipartite levels matching so we have a closer set to look through and we find that much like for 1930, there are much fewer network connections for women than for men suggesting that the noun-verb connections has not really improved over time. The exact words that show up are slightly different with kiss, understand and look showing up prominently for women compared to more prominent actionable terms for men like fight,

choose, grow and hurt as discussed before. This action bias towards males becomes even clearer when we look at the noun-adjective relationships.

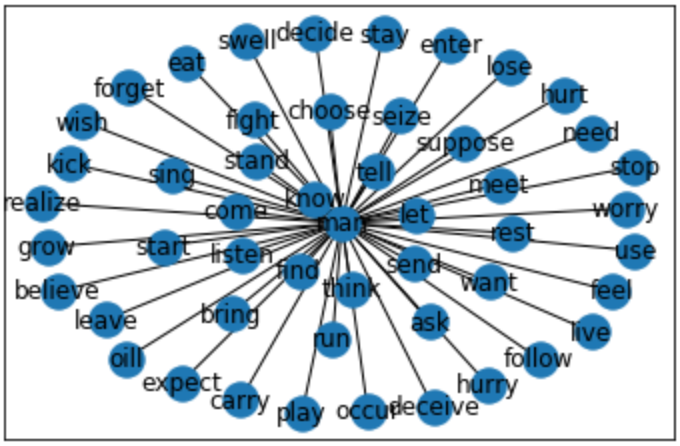
Noun-Verb network for Woman
(1930)



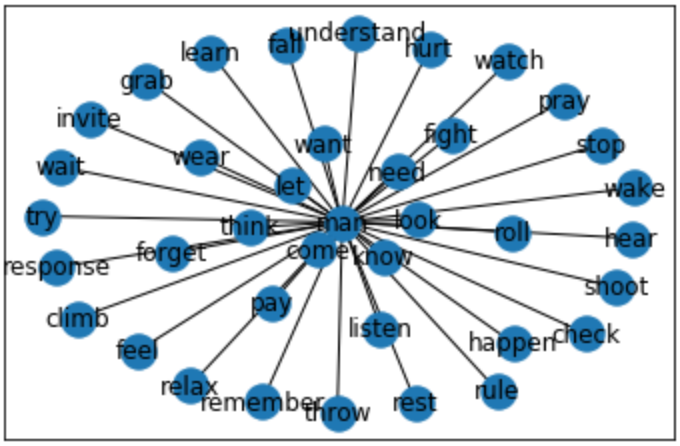
(2010)



Noun-Verb network for Man
(1930)



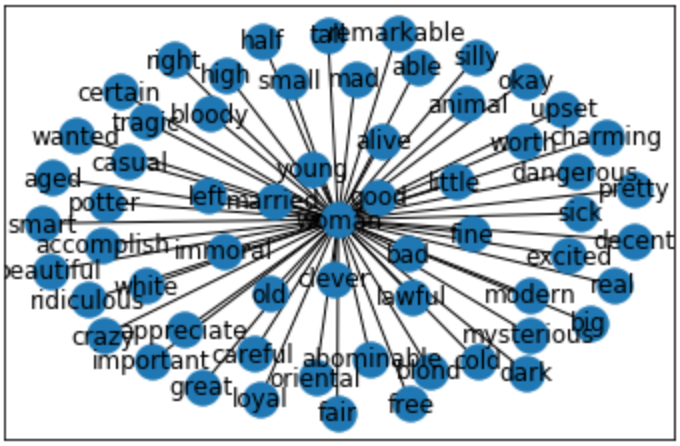
(2010)



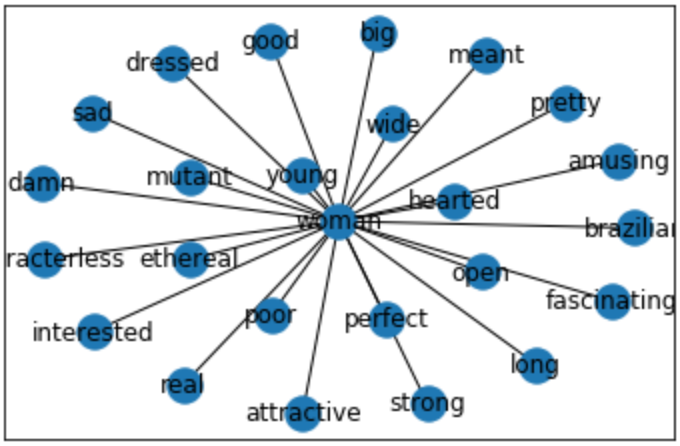
Noun-adjective relationships show a slightly different story for gender relations with 1930 woman relations having words like married and lawful quite close to the center however on the periphery, we do see diverse descriptive terms ranging from more appearance related descriptors like pretty and beautiful to words that to me carry more heft when giving a character a background and development such as strong, abominable, etc. but these are also peppered among other words like silly and crazy which do fall into common female character movie tropes, portraying women as a crazy -ex girlfriend or a blonde ditz for example. When we use the degree and eigenvector centralities to sort the centrality of these terms, we do get the same words which include great, charming, crazy and alive so different aspects of personality. 2010 adjective similarities are similar in this respect with what I thought were diverse descriptors on the whole. The men as expected in both the 1930s and 2010s are much more saturated in their descriptors despite the same node weight restrictions, once again implying their dominance in the film scripts in both magnitude and depth of characters they portray. I query these time-related differences further in the next section using vector space relations to see if we can get clearer

distinctions between men and women as well as between women of different time periods and genres as well.

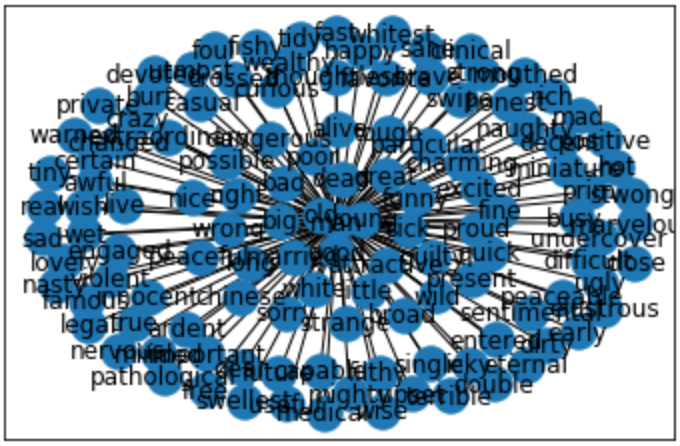
Noun-Adjective network for Woman
(1930)



(2010)



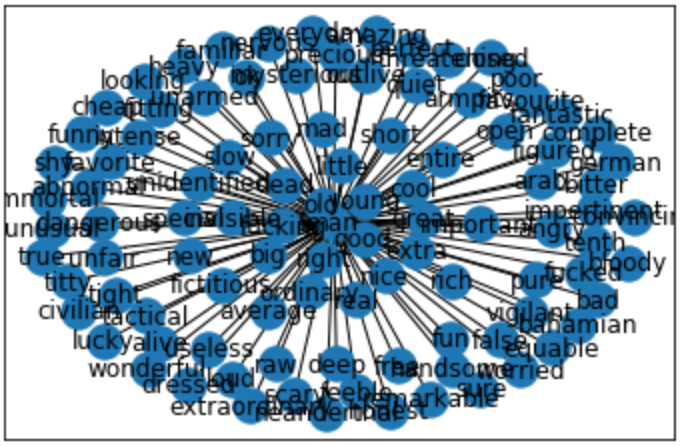
Noun-Adjective network for Man
(1930)



```
sorted(nx.eigenvector_centrality(g_manNJ30).items(), key = lambda x: x[1], reverse = True)[:10]
```

```
[('man', 0.7071109838196059),
 ('marvelous', 0.06428205259350105),
 ('ugly', 0.06428205259350105),
 ('crazy', 0.06428205259350105),
 ('alive', 0.06428205259350105),
 ('wet', 0.06428205259350105),
 ('positive', 0.06428205259350105),
 ('minded', 0.06428205259350105),
 ('favorite', 0.06428205259350105),
 ('clinical', 0.06428205259350105)]
```

(2010)



```
sorted(nx.eigenvector_centrality(g_manNJ10).items(), key = lambda x: x[1], reverse = True)[:10]
```

```
[('man', 0.7071107072015742),
 ('bitter', 0.07106651087032304),
 ('alive', 0.07106651087032304),
 ('favorite', 0.07106651087032304),
 ('true', 0.07106651087032304),
 ('sorry', 0.07106651087032304),
 ('cheap', 0.07106651087032304),
 ('worried', 0.07106651087032304),
 ('perfect', 0.07106651087032304),
 ('fantastic', 0.07106651087032304)]
```

Vector Space Relations

Most Similar Words and Documents

Researching the most similar words and documents throughout the years for all the movies without even any genre or decade isolations show some very revealing information about how men and women are written about in movies. I'm using the doc2vec relationships over word2vec as the document relationships are going to provide more information about how gender is situated in documents as a whole rather than looking at specific word relationships which we look more into when we do the parts of speech analysis. Firstly, from the doc2vector under unsupervised modeling through gensim, we establish tagged documents for both female and male documents by gender.

Firstly, when we look at gender relationships using documents that were tagged with 'female' related words such as woman, girl, etc. in the context of all 9909 movies regardless of year and genre, we find some interesting, quite sexist words that are very telling of how female characters are usually written about and discussed within movies. The descriptors

of their hair and appearance (blonde, brunette, redhead) as well as both words like minx and on the other side of the spectrum words like mousy are revealed as the most similar words when women and girls are the words compared. This sexism is also echoed when we look into the documents tagged with male keywords where the words are also hair colour descriptors including brunette and blonde in addition to words like bombshell. These words also have quite high similarity values, mostly above 0.5 for the women.

On the other hand, for male keywords within female and male tagged documents, we see that the similar words tend to just be other colloquialisms and words used to call other men, such as buddy, homie, homeboy, lad, fam, for example, in both male and female tagged documents. It is important to note that these words might appear in the male related word similarities due to the fact that 'man' is a sort of general term to address people so words like buddy, homie, bro align very closely to that definition. Nonetheless, these similarities reveal on average, movies use very stereotypically descriptive and sexist terminology to describe women, and these stereotypes do not seem to affect the male gender as much. We query this even further using genre and time-related breakdowns below.

Female Similar Words - All movies

Documents tagged with female keywords

```
mvD2V_female.most_similar(positive = ['woman','girl'], negative = ['man'], topn = 10)
```

```
[('brunette', 0.5521952509880066),
 ('blonde', 0.547403872013092),
 ('redhead', 0.5351743698120117),
 ('prettiest', 0.5166902542114258),
 ('pinup', 0.5048144459724426),
 ('flirt', 0.4958895742893219),
 ('minder', 0.4949256181716919),
 ('niece', 0.489028662443161),
 ('bombshells', 0.48542696237564087),
 ('blond', 0.48342978954315186)]
```

Documents tagged with male keywords

```
mvD2V_male.most_similar(positive = ['woman','girl'], negative = ['man'], topn = 10)
```

```
[('redhead', 0.5722601413726807),
 ('brunette', 0.5712414979934692),
 ('blonde', 0.565335750579834),
 ('mousy', 0.5193119049072266),
 ('unisex', 0.5185565948486328),
 ('bombshells', 0.5135267972946167),
 ('prettiest', 0.5078239440917969),
 ('daughter', 0.49182310700416565),
 ('minx', 0.4886482357978821),
 ('carla', 0.48468464612960815)]
```

Male Similar Words - All movies

Documents tagged with female keywords

```
mvD2V_female.most_similar(positive = ['man','boy'], negative = ['woman'], topn = 10)
```

```
[('buddy', 0.5500340461730957),
 ('rascals', 0.5384502410888672),
 ('homie', 0.5353148579597473),
 ('dawg', 0.5141319036483765),
 ('lad', 0.502248227596283),
 ('homies', 0.5002907514572144),
 ('marky', 0.4982272684574127),
 ('homey', 0.4966511130332947),
 ('fam', 0.49355340003967285),
 ('homeboy', 0.49234485626220703)]
```

Documents tagged with male keywords

```
mvD2V_male.most_similar(positive = ['man','boy'], negative = ['woman'], topn = 10)
```

```
[('buddy', 0.5860825181007385),
 ('atta', 0.5367481708526611),
 ('jangers', 0.5269747972488403),
 ('jamesy', 0.5155718922615051),
 ('lad', 0.5068671703338623),
 ('homie', 0.5006662607192993),
 ('scout', 0.48949190974235535),
 ('dink', 0.4848521053791046),
 ('bro', 0.48441779613494873),
 ('drummer', 0.4813867509365082)]
```

Now, looking specifically at time related differences, we find that from 1930 to 2010, we do see that the terminology similarities have changed to some extent.

At first glance, it appears that the word associations from the 1950s are less misogynistic in nature given the reduced number of appearance related descriptors like those that appear for the 2010s word associations with prettiest, blonde, cutest, etc. Looking at the 1950s word associations, the first two words that appear are 'faithful' and 'sophisticated', both with high word similarity values at around 0.84 for women and other words in the 10 ten most similar include 'modest' and 'marriage'. These words do conform to the conventional submissive stereotype that is associated with females from an earlier time period and particularly the 1950s, reflecting the post-war gender stereotypes. Postwar prosperity made the banalities of housework less taxing, but often came at a cost to women who gave up careers to maintain the domestic sphere. This lifestyle stressed the importance of a one-income household, with the husband working and the wife staying at home to raise the children which seems to be reflected in the word usage of the time or at least in the movies that were released at the time. These home-bound terminologies seem to have evolved into more explicit terms related to appearance as seen in the movies of the 2010s with words like 'attractive', 'sexy' and 'passionately' for example as seen when we look at words

associated with women and girls in female tagged documents. Male tagged movies too have similar comparisons with the 1950s movies in this case with adverbs like sweetly and adjectives like inquisitive and nouns like tenderness contrasted with once again more explicit terms like orgasm and wetting for 2010.

Female Similar Words - 1950 movies compared to 2010 movies

```
mvD2V_female50.most_similar(positive = ['woman','girl'], negative = ['boy', 'man'], topn = 10)
```

```
[('faithful', 0.8399957418441772),
 ('sophisticated', 0.8392194509506226),
 ('attractive', 0.8286294937133789),
 ('mustache', 0.8201954364776611),
 ('unreal', 0.8033789396286011),
 ('marriage', 0.7966414093971252),
 ('quality', 0.7872207164764404),
 ('modest', 0.7836649417877197),
 ('farthing', 0.7808701992034912),
 ('women', 0.7785411477088928)]
```

```
mvD2V_female10.most_similar(positive = ['woman','girl'], negative = ['boy', 'man'], topn = 10)
```

```
[('attractive', 0.5501745939254761),
 ('lady', 0.5045493841171265),
 ('sexy', 0.45774829387664795),
 ('women', 0.45704615116119385),
 ('virtuous', 0.4526621103286743),
 ('lesbian', 0.4518219232559204),
 ('vagina', 0.43385642766952515),
 ('succubus', 0.43176722526550293),
 ('extensions', 0.43111667037010193),
 ('passionately', 0.4297320246696472)]
```

Now looking at genre specific comparisons, we do not really find extremely revealing word vector insights that are explicit to the genre being analyzed. The words seem to be more related to general themes and objects reflective to the genre itself not really descriptive of the men and women descriptors or verb relations like we had seen more clearly in the decade comparisons. Particular genres are more explicitly revealing such as romance and comedy films, where we see words like pretty, attractive and prostitute, flirt and other

blonde for example. For other more distinct genres like sport, horror or sci-fi, we do not get many terms that are interpretable; I tested different female and male word combinations for the different genres to check perhaps whether words like woman and girl were not being used in the first place itself, instead testing words like 'wife' and 'daughter'. Nonetheless, looking at the word similarities that show up, we can maybe make some interesting conclusions : Words like orphan, child, sister and angel within horror films might reveal the type of characters that females typically play within this genre. I then decided to look more into the word embedding relationships to further extract how gender is situated within the genres specifically.

```
mvD2V_rom_female.most_similar(positive = ['woman','girl'], negative = ['boy'], topn = 10)
```

```
[('attractive', 0.543552041053772),
 ('desirable', 0.5255258083343506),
 ('prostitute', 0.5176061391830444),
 ('vivacious', 0.5054903030395508),
 ('impressionable', 0.503525972366333),
 ('ideal', 0.4975433051586151),
 ('magda', 0.49329113960266113),
 ('adele', 0.4838694930076599),
 ('demure', 0.48195120692253113),
 ('intelligent', 0.47508466243743896)]
```

```
mvD2V_com_female.most_similar(positive = ['woman','girl'], negative = ['man', 'boy'], topn = 10)
```

```
[('redhead', 0.5138537883758545),
 ('peignoir', 0.5102139711380005),
 ('prettiest', 0.5037169456481934),
 ('attractive', 0.5025854110717773),
 ('scrubba', 0.492868572473526),
 ('charmer', 0.4899721145629883),
 ('blonde', 0.47883540391921997),
 ('flirt', 0.46782612800598145),
 ('daughter', 0.46521151065826416),
 ('schoolgirl', 0.4647250175476074)]
```

Word Embeddings

From word embedding relationships, we can clearly see that war movies are more heavily skewed to the male gender and family and musicals are more related to women. When we try and map some intersectionality within the films using occupation, genre and class we can see the types of movies that are associated with one race versus other types of movies which is quite interesting and also suggests that movie studios should make changes to the stereotypical way they portray certain genders and certain races. The word cloud below further corroborates this sentiment with the different sizings of the word clouds. One could make the argument for gender specifically, that women like to see certain types of films so are more likely to be portrayed in some genres more than others but this same argument cannot really be made for races. My research focuses just on gender differences owing to time and computational constraints but more research on the racial diversity must be conducted especially in the current heated climate.

Conclusion

Overall, it is evident that women and men are written about quite differently in terms of their autonomy as characters as well as their relationships with other characters as can be seen by the most common verbs and adjectives related with their existence in films. Certain genres are also more likely to cater to women compared to others. Similarly, there are some temporal changes that can be observed when describing male and female characters, however these changes are not necessarily unequivocally positive. Women are written about in a less stereotypically submissive lens but this seems to have just 'evolved' into more misogynistic, sexist descriptions of their appearance which I would not classify as change in the most positive sense. Given that, recent evidence has revealed that female-led and movies that were more gender inclusive do actually perform well at the box office. This

strongly suggests that script-writers need to alter the way in which women are described and written about both linguistically as well as accounting for their position and role within the greater context of the film. Some ways of making these improvements could be hiring a more diverse production cast and crew and more female and/or directors of color to helm more gender, race and ideology inclusive films to help influence social norms for the better. This could go a long way in changing the way women are portrayed in film which could in turn lead to a positive attitude change in the general public and in gender norms as a whole over time hopefully.